IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Gaze-Informed Egocentric Action Recognition for Memory Aid Systems

ZHEMING ZUO[1], LONGZHI YANG[1], (Senior Member, IEEE), YONGHONG PENG[2], (Member, IEEE), FEI CHAO[3], (Member, IEEE), YANPENG QU[4]

[1]Department of Computer and Information Sciences, Northumbria University, NE1 8ST, U.K. (e-mails: {zheming.zuo, longzhi.yang}@northumbria.ac.uk)
[2]Faculty of Computer Science, University of Sunderland, St Peters Campus, Sunderland, SR6 0DD UK. (email: Yonghong.Peng@sunderland.ac.uk)
[3]Cognitive Science Department, Xiamen University, P. R. China (email: fchao@xmu.edu.cn)
[4]Information Science and Technology College, Dalian Maritime University, P. R. China (email: yanpengqu@dlmu.edu.cn)

Corresponding author: Longzhi Yang (e-mail: longzhi.yang@northumbria.ac.uk).

This work was supported by the National Natural Science Foundation of China (No. 61502068).

ABSTRACT Egocentric action recognition has been intensively studied in the fields of computer vision and clinical science with applications in pervasive health-care. The majority of the existing egocentric action recognition techniques utilize the features extracted from either the entire contents or the regions of interest in video frames as the inputs of action classifiers. The former might suffer from moving backgrounds or irrelevant foregrounds usually associated with egocentric action videos, while the latter may be impaired by the mismatch between the calculated and the ground truth regions of interest. This paper proposes a new gaze-informed feature extraction approach, by which the features are extracted from the regions around the gaze points and thus representing the genuine regions of interest from a first person of view. The activity of daily life can then be classified based only on the identified regions using the extracted gaze-informed features. The proposed approach has been further applied to a memory support system for people with poor memory, such as those with Amnesia or dementia, and their carers. The experimental results demonstrate the efficacy of the proposed approach in egocentric action recognition and thus the potential of the memory support tool in health care.

INDEX TERMS Gaze-informed egocentric action recognition, gaze-informed region of interest, UNN-GazeEAR data set, memory aid.

I. INTRODUCTION

VISUAL human action recognition is a growing research field applicable to areas including human computer interaction, video surveillance, motion analysis, and recently clinical science. The focus of such research usually lies in the classification of movement patterns with reasonable or near perfect performance demonstrated on the common benchmark data sets [1]–[3]. Different with the scenarios included in these data sets, people in the real world interact with other people as a natural part of performing a daily activity, which are witnessed and perceived by the actors from their own egocentric point of view; such egocentric view of interactions are important part of the visual evidence that should be considered when performing action recognition tasks. Also, it is very challenging to capture all the interactive activities of a person using static or smoothly moving cameras.

An alternative to the conventional “third person" video capturing, to address the challenges, is the use of a portable or wearable video capturing device, often a pair of smart glasses, to record all the activities of a person from an egocentric perspective or a first person point of view [4]. The egocentric videos, in contrast to the conventional ones, can usually minimize occlusions and often present the interaction activities in the centers of video frames. Various conventional action recognition approaches have been directly applied to the egocentric action recognition approaches as reported in the literature. For instance, the optical flow-based motion descriptors were used to discover ego-actions in sport videos [5]; and a low-dimensional signature vector GIST was applied to egocentric videos for first-person video analysis in the work of [6]. Moving backgrounds are naturally associated with egocentric videos, which significantly affect the effectiveness of the conventional action recognition technologies.

Region-based action recognition approaches can be used to address the issue of fast moving background. By noticing that the regions of interest (ROI) is usually appeared in the vicinity of hands [7], [8], regions around hands have thus been used as the ROIs [9]. However, the effects of this type of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808486, IEEE Access

approaches may be limited by the partial occlusion of hands, which often happens in egocentric videos. Saliency maps were proposed for ROI determination based on the fact that the most important activity information is usually appeared in the foregrounds [8]. The image signature was one of the most effective saliency map generation approaches, which was proposed as the sign function of the Discrete Cosine Transform (DCT) of an image, followed by the inverse DCT operation [10]. Saliency maps-based ROI detects the location of ROI by mimicking the process of human eye fixation in egocentric vision. The performance of this approach decreases when the approach is applied to egocentric videos with large camera motion, the effect of which usually can be compensated by a person [11].

This paper proposes a new gaze-informed egocentric action recognition approach to enhance the saliency-based method, given that gaze capturing function has been integrated into more and more smart glasses. Noting that multiple saliency regions are often resulted from saliency map-based ROI for a given threshold value, the proposed approach singles out the region where the gaze is located and performs egocentric action recognition based on this selected region only, as illustrated in Fig. 1. The proposed gaze-informed ROI (GROI) not only benefits from the advantage of general ROI-based action recognition approaches with higher computational efficiency, but more importantly also suppresses the interference of moving backgrounds and other irrelevant foregrounds.

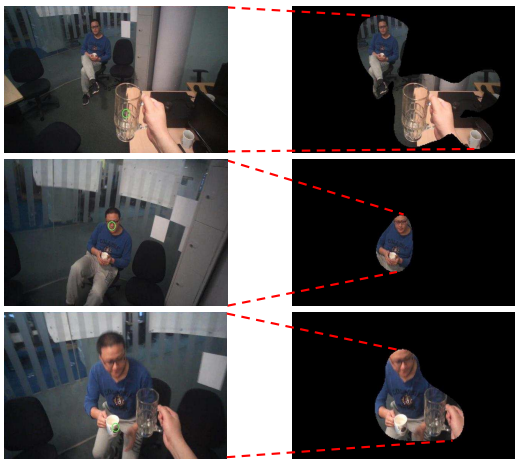


FIGURE 1: Illustrative original frames with gaze data (green circles) shown on the left and the corresponding gaze-informed ROIs on the right

This paper also presents a memory aid system developed upon the proposed gaze-informed egocentric action recognition approach. Using this system, the patients with poor memory and their carers can make queries of the patients' important activities in daily life (ADLs), and five typical human to human interactive ADLs have been utilized in this work for system validation and evaluation. The contribution of this paper is threefold: 1) proposing a gaze-informed ROI determination method and an egocentric action recognition framework using the proposed region determination method and the bag of visual words (BoVW) for local feature extraction, 2) capturing an ADL data set UNN-GazeEAR[†], and 3) developing a memory aid system which takes text/speech queries of ADLs and returns the corresponding video clips for the queried ADLs. The proposed system is validated and evaluated using the captured UNN-GazeEAR data set with comparative results generated.

The rest of the paper is organized as follows. Section II reviews the technical background and related work. Section III details the proposed gaze-informed egocentric action recognition system. Section IV presents the memory aid system for people with poor memory. Section V reports the experiments for system validation and evaluation. Section VI concludes this work and points out future research directions.

II. BACKGROUND

ROI detection and background removal are the two most common pre-processing approaches before the application of action classifiers. In particular, ROI can be any candidate regions under consideration in a single image or video frame. Deep learning approaches, such as Region-based Convolutional Neural Network (RCNN) and fast RCNN, have been applied for feature map (i.e., a number of high scoring ROIs) generation [12]. These models are powerful but often time consuming. Saliency detection aims to quantitatively represent how human eyes identify important objects/people in a scene. Therefore, ROI may be predicted using a saliency map [11], which is generally considered as a cheap method for detecting ROIs. For instance, saliency was considered as a problem of small (in terms of spatial sparsity) foreground in a simple (in terms of spectral sparsity) background [10]. This leads to the proposition of the computationally cheap image signature, which is a simple yet efficient extension of the spectral saliency to approximate or predict the foreground spatial location that may be consistent with the human eye movement fixations.

Once the ROIs for all the input videos are detected, feature extraction is applied to the ROIs, which transforms the ROIs into a set of features that well represent the actions of the videos. Feature extraction can be implemented locally or globally. Typical local video representations can be built using techniques such as space time interest points (STIP) [13], histogram of oriented gradients (HOG) [14], histogram of optical flow (HOF) [15], motion boundary histogram (MBH) [16], and histogram of motion gradients (HMG) [17], [18]. These feature descriptors have been widely used to capture and analyze the local properties of the visual spatio-temporal video streams. The improved dense trajectories (IDT) representation combines HOG, HOF, and MBH, which has been applied to a collection of challenging action data sets [19]. IDT can also be jointly utilized with

[†]The UNN-GazeEAR data set is available online at: www.lyang.uk/UNN5-GazeEAR.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808486, IEEE Access

other feature descriptors for performance enhancement by means of early or late fusion [17].

To further advance the compactness and discriminative ability of the locally extracted features, feature encoding is often employed. Feature encoding aims to generate uniform visual representations (typically for a set of videos and each of which is of different time length) that commonly implemented in the form of single fixed-length visual words encoded under the BoVW paradigm. Early encoding methods transform a set of local descriptors into a single vector of visual words using cluster algorithms, such as k-means [20]. Until recently, super-vector based encoding schemes have been widely employed in the tasks of action recognition. By aggregating high order statistic values, this sort of encoding schemes usually result in feature representations with very high dimensionality. Typical super-vector based encoding methods include local tangent-based coding, super vector coding, vector of locally aggregated descriptors (VLAD), and fisher vector (FV) [21]. In particular, compared with FV, VLAD is generally practically faster but may suffer from relatively poor performance. This is because VLAD a simplified non-probabilistic version of FV and it is not able to capture spatial information from the extracted features [21].

The encoded features are the inputs of classifiers for action recognition tasks. The most widely used classifiers include artificial neural networks (ANNs) [22] and support vector machine (SVM) [17], [19]. Interestingly, linear [17], [19], rbf-χ^2 kernel [19], or histogram intersection kernel SVMs is often coupled with spatial pyramid and dense sampling strategy [23], (sum or mean) pooling strategies, and normalization (such as $\ell 2$ or power normalization plus $\ell 2$) to perform non-linear action classification tasks for good recognition accuracy.

III. GAZE-INFORMED EGOCENTRIC ACTION RECOGNITION

The framework of the proposed gaze-informed egocentric action recognition (EAR) approach is illustrated in Fig. 2, which integrates the gaze-informed ROI (GROI) detection into the BoVW framework. The approach mainly consists of four phases, including GROI detection, feature extraction, feature encoding, and classification. GROI identifies the regions directly related to the actions implied by the video clips for processing, which helps in suppressing the impact of noisy backgrounds and irrelevant foregrounds information. Based on the determined GROIs, feature extraction obtains discriminatory information from image sequences that is robust in distinguishing the ADLs. From this, the extracted feature values of each ADL are encoded into an identifier of the particular activity. Finally, a trained classifier is employed to recognize the actions expressed in egocentric video clips using the identifier.

It is a common practice to apply the pre-processing and post-processing techniques before and after the feature encoding phase, for improving the performance of action recognition, which are also applied in this work as illustrated in Fig. 2. Note that a good range of general purpose classifiers, such as ANN and SVM, can be readily employed in the final classification phase, and thus this phase is not detailed in this section while others are detailed below.

A. GROI DETECTION

The ROIs for a given video clip regarding an ADL can be represented as the collection of ROIs in all video frames, as a video clip is essentially a series of frames or images. The ROI in an image has been intensively studied in the literature. Generally, an image can be expressed as multiple foreground regions and a background. The ROI is usually assumed as a foreground region in the literature, which is also the assumption in this work. Saliencies have been widely used in ROI detection in the field of image processing and computer vision. Note that multiple saliency regions are often detected using the current saliency detection approaches for a given threshold, while only one ROI is really associated to each frame of each ADL. This work therefore further develops the saliency-based ROI detection approach by taking in the gaze information, given that gaze is one of the most significant types information in indicating ROIs in first-person videos. To facilitate the discussion, a video clip in this paper is represented as a series of frames $\{\mathcal{F}r_1, \mathcal{F}r_2, \cdots, \mathcal{F}r_f\}$ along the time line, each with resolution of $r_x * r_y$. As the same GROI detection operation is applicable to every frame, a random frame $\mathcal{F}r_i, 1 \leq i \leq f$, is taken as an example, without losing the generality, in introducing the GROI detection approach in the remainder of this sub-section.

1) Representation space transformation

In order to generate the saliency map of frame $\mathcal{F}r_i$, its RGB representation from the camera viewpoint is mapped to the LAB space denoted as $\mathcal{F}_i$, as shown in Fig. 2. Similar to RGB, LAB also consists of three channels, including brightness in the range of [0,100], redness-greenish in the range of [-100, 100] where positive values represent redness and negative ones represent greenish, and yellowish-bluish in the range of [-100, 100] where positive values indicate yellowish and negative ones represent bluish. Despite of the same dimensions, RGB is a linear space while LAB is a non-linear one. RGB representation is most commonly employed in our daily life as it is less abstract compared to LAB, but LAB is usually a better choice in the egocentric visual processing due to the non-linearity of human perception in colors [24].

2) Saliency map generation

The saliency map is fundamentally a spatial support of foreground (i.e., a set of non-zero elements) in an LAB-typed egocentric video frame $\mathcal{F}_i$. This can be approximately isolated by taking the $sign$ operation of the mixed signal $\mathcal{F}_i$ in LAB representation in the transformed space through a discrete cosine transform (DCT), followed by an inversely transform which takes it back to the spatial domain using the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808486, IEEE Access

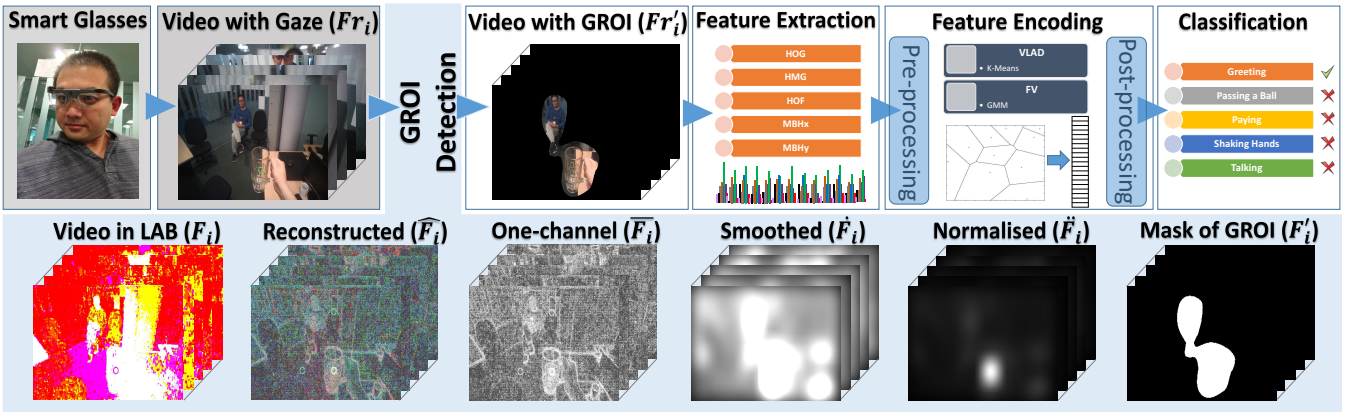


FIGURE 2: The framework of gaze-informed egocentric action recognition, and the details of the main component GROI detection

inverse DCT (IDCT) operation [10]. Formally, given an LAB frame $\mathcal{F}_i$, the reconstructed frame $\hat{\mathcal{F}}_i$ is defined by:

$$\hat{\mathcal{F}}_i = IDCT[sign(\tilde{\mathcal{F}}_i)], \tag{1}$$

where $\tilde{\mathcal{F}}_i = DCT(\mathcal{F}_i)$, $sign(\cdot)$ is the entry-wise operation for isolating the support in the foreground of the frame in the transformed space, $IDCT(\cdot)$ represents the operation of inverse transformation that converts the results back into the spatial space.

Once the reconstructed frame $\hat{\mathcal{F}}_i$ is obtained, the three channels of $\hat{\mathcal{F}}_i$ is integrated into a one-channel frame $\bar{\mathcal{F}}_i$ by averaging the pixel values in the L, A, B channels:

$$\bar{\mathcal{F}}_i = \frac{1}{3}(\hat{\mathcal{F}}_{iL} + \hat{\mathcal{F}}_{iA} + \hat{\mathcal{F}}_{iB}), \tag{2}$$

where $\hat{\mathcal{F}}_{iL}$, $\hat{\mathcal{F}}_{iA}$, and $\hat{\mathcal{F}}_{iB}$ represent the frame values in L, A, and B channels, respectively. Then, the one-channel frame $\bar{\mathcal{F}}_i$ is smoothed by:

$$\dot{\mathcal{F}}_i = \kappa_g * (\bar{\mathcal{F}}_i \circ \bar{\mathcal{F}}_i), \tag{3}$$

where $(\circ)$ denotes the entry-wise Hadamard product operations and κ_g represents the Gaussian kernel. From this, the final saliency map $\ddot{\mathcal{F}}_i$ is obtained by normalizing $\dot{\mathcal{F}}_i$ to the range of [0, 1], where 0 represents black and 1 represents white in Fig. 2. Note that the computation of the frame-wise saliency map $\ddot{\mathcal{F}}_i$ can be speeded up by resizing the resolution of $\mathcal{F}r_i$ with a scaling factor Ω. For instance, $\Omega = 6$ is applied in the experimentation as detailed in Section V-B.

3) GROI Determination

A saliency map represents the probabilities of different regions being the real ROI in an image, but the region with the highest saliency value may not be consistent with that of the actor in a video clip. For instance, there are a group of people talking to each other in a frame, while the actor is looking at and talking to one person who does not belong to the group and appears in a different location in the frame. In this situation, the saliency map may highlight the group of people, but the real ROI is the individual person as indicated by the gaze point. Notice that the gaze point is usually associated with the real region of interests. Therefore, the gaze information is introduced in the ROI detection process in this work as summarized in Algorithm 1.

Algorithm 1: Gaze-informed ROI detection

Input : $\ddot{\mathcal{F}}_i$,: saliency map of the ith frame
(x_i, y_i): the coordinate of the gaze point
τ: default threshold for GROI detection
ϵ: GROI mask expansion width

Output: $\mathcal{F}'_i$: mask of identified GROI

1: Initialize τ, ϵ, and $\mathcal{F}'_i(x_j, y_k) \leftarrow 0$ where $1 \le x_j \le r_x$ and $1 \le y_k \le r_y$;
2: **if** $\ddot{\mathcal{F}}_i(x_i, y_i) > \tau + \epsilon$ **then**
3: **return** $getRegion(\mathcal{F}'_i, \ddot{\mathcal{F}}_i, (x_i, y_i), \tau + \epsilon)$;
4: **else**
5: $\tau' = \ddot{\mathcal{F}}_i(x_i, y_i) - \epsilon$;
6: **return** $getRegion(\mathcal{F}'_i, \ddot{\mathcal{F}}_i, (x_i, y_i), \tau')$;
7: **end if**

The algorithm takes four inputs, including the saliency map $\ddot{\mathcal{F}}_i$, the gaze point coordinate (x_i, y_i), the default saliency threshold for ROI τ, and the saliency map expansion width ϵ. The default saliency threshold for ROI is usually set as 0.5, based on the fact that the foreground is often a small part of the image and the assumption that the ROI is a foreground region as discussed earlier. The GROI mask expansion width ensures there is a margin between the gaze point and the edge of the GROI region. The value of ϵ is empirically determined. If the saliency value of the gaze point is greater than the threshold including the margin, the algorithm extracts the saliency region which includes the gaze point as expressed in Lines 2 and 3; otherwise, the threshold is adaptively modified to make sure the gaze point is included in the saliency region based on the modified threshold, as described in Lines 5 and 6.

The sub-procedure $getRegion$, as shown in Algorithm 2,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808486, IEEE Access

extracts the GROI region by producing a mask frame where the region points are valued as 1 and all the other points are valued as 0. The algorithm firstly checks if a given point is within the range of the frame. Note that, as introduced in the beginning of this section, the resolution of the frame is $r_i * r_y$ If the point is within the range of the frame and its saliency value is greater than the threshold, the mask value of this point will be updated as 1; otherwise the value will be unchanged as its default value of 0. As only one connected region that includes the gaze point needs to be identified, the sub-procedure starts from the gaze point and recursively examines its surrounded 4 neighboring points, until the algorithm reaches the frame boundary or the saliency value of the currently examined point is less than the threshold.

Algorithm 2: Region extraction procedure

Input : $\mathcal{F}r'_i$: current GROI mask frame
$\ddot{\mathcal{F}}_i$,: saliency map of the ith frame
(x_i, y_i): coordinate of gaze point in ith frame
τ: determined saliency threshold for ROI

Output: $\mathcal{F}r'_i$: mask of identified GROI

1 **Procedure** $getRegion(\mathcal{F}'_i, \ddot{\mathcal{F}}_i, (x_i, y_i), \tau)$

1: **if** $1 \leq x_i \leq r_x$ and $1 \leq y_i \leq r_y$ and $\ddot{\mathcal{F}}_i(x_i, y_i) \geq \tau$ **then**
2: $\mathcal{F}'_i(x_i, y_i) \leftarrow 1$;
3: $\mathcal{F}r'_i = getRegion(\mathcal{F}'_i, \ddot{\mathcal{F}}_i, (x_i - 1, y_i), \tau)$;
4: $\mathcal{F}r'_i = getRegion(\mathcal{F}'_i, \ddot{\mathcal{F}}_i, (x_i + 1, y_i), \tau)$;
5: $\mathcal{F}r'_i = getRegion(\mathcal{F}'_i, \ddot{\mathcal{F}}_i, (x_i, y_i - 1), \tau)$;
6: $\mathcal{F}r'_i = getRegion(\mathcal{F}'_i, \ddot{\mathcal{F}}_i, (x_i, y_i + 1), \tau)$;
7: **end if**
8: **return** $\mathcal{F}'_i$;

Note that the computational efficiency of Algorithm 2 may be improved by reducing the number of repeated examinations for the same points, but this is not considered in the pseudocode for the purpose of better readability. The final output of Algorithm 1 is a mask frame which masks all the points within the identified GROI with value 1 and those outside the range as 0. From this, the mask is applied to the original video frame $\mathcal{F}r_i$ in RGB representation using a simple multiplication operator, and then the GROI of the frame $\mathcal{F}r'_i$ can be generated for further processing as shown in Fig. 2.

B. FEATURE EXTRACTION

After the GROI is generated, feature extraction approaches need to be applied to distinguishably represent the video actions. In particular, the local feature descriptors HOG [14], HOF [15], MBH [16] and HMG [17] are applied in this work. Conventionally, a video clip can be represented as a three-dimensional volume. The local descriptors of each pixel of each frame is firstly calculated using gradients (HOG and HMG), optical flow (HOF) or their combination (MBH). Then, the calculated descriptor is evenly divided into a predefined number of spatio-temporal blocks, in order to speed up the feature extraction process and improve its generalization ability [25]. After this, the orientations of gradient or optical flow of each block are summarized as a partial histogram, and all such partial histograms from all blocks are finally concatenated into a single feature vector representing the action in the video.

HOF represents motions using the optical flow $\vec{OF}$, that is the appearance of brightness patterns. A number of approaches have been proposed for the calculation of optical flow, and the Horn-Schunk method is particularly adopted in this work due to its comparative performance [26]. HOG uses the spatial gradients as the basic descriptors of local features, which is fundamentally a derivative operation but almost always practically implemented using a convolution operation with various kernels for fast processing. HMG is a direct extension of HOG by considering the temporal information before the spatial gradient calculation through a derivative operation on each pair of neighboring frames. The MBH combines the optical flow and gradient operations, which takes the horizontal and vertical components of the optical flow resulted from HOF separately followed by the gradient calculation as used in HOG and HMG.

The resulted optical flow from HOF and gradient information from other descriptors are all vectors, expressing the magnitude and direction of such information of each pixel. From this, the orientations in the two dimensional frame space are evenly quantized into a number of bins (usually valued as 8) in the range of [0, 2π], and each pixel is partially assigned to the two closest bins that flank the gradient orientation (or one bin when the bin and the gradient direction match exactly) based on the bi-linear interpolation. The strength of each bin (represented as one bar in a partial histogram expressing the block) is then calculated as the weighted summation (or its variation [27]) of the magnitudes of its assigned pixels. Finally, the partial histograms led by different blocks are concatenated into a single feature vector which is the visual representation for the action implied by the video.

C. FEATURE ENCODING

1) Pre-processing

The local features are usually of high dimensionality and a high degree of correlation. This makes the subsequent unsupervised codebook generation difficult, which usually implemented using k-means and Gaussian mixture model (GMM) clustering, in the feature encoding phase. To address this, pre-processing phase is often applied, which consists of feature normalization and dimensionality reduction. One of the most commonly used feature normalization approaches is the RootSIFT [28], which is also used in this work. RootSIFT takes the element-wise square root of the $\ell 1$ normalization. This operation utilizes the Hellinger distance instead of the commonly used Euclidean distance as the latter is often dominated by large bin values and the former is also sensitive to smaller ones.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808486, IEEE Access

There is a large number of dimensionality reduction techniques that can be readily applied for action recognition [29]. Among them, the principle component analysis (PCA) is the most widely used unsupervised feature reduction techniques, which is also adopted in this work. It can be regarded as a statistical procedure to process the extracted local features by projecting the map feature to a set of linearly uncorrelated variables (i.e., principle components) using the orthogonal transformation. The typical number of principle components is less or much less than the number of variables in the original extracted local features, it is therefore resulting in effective dimensionality reduction.

2) The Encoding

Visual words encoding under the paradigm of BoVW further reduces the dimensionality and addresses the possible indexability issue of the local features. In particular, two supervector based encoding methods VLAD [30] and FV [31] are used in this work. VLAD starts with the calculation of centroids in the feature space using k-means clustering algorithm, which is followed by the aggregation of features using the calculated centroids. FV firstly computes the Gaussian mixture model (GMM) with $2nd$ order information, which is then used to aggregate the local features based on its mean and covariance. These two encoding methods result in encoded features with different dimensions with k-by-d from VLAD and 2-by-k-by-d from FV, where k is the number of centroids (i.e., visual words) and d is the dimensionality of the encoded feature vector. Indeed, VLAD is a simplified non-probabilistic version of FV, and thus VLAD is practically faster but with relatively poorer performance.

The codebook generation methods used in VLAD [30] is the k-means. Given a set of preprocessed local features $S = \{s_1, s_2, \cdots, s_n\}$, where n is the number of local features (i.e., the number of videos in visual action recognition) and s_i $(1 \leq i \leq n)$ denotes the ith local feature, suppose k clusters $\{c_1, c_2, \cdots, c_k\}$ is required, where c_j $(1 \leq j \leq k)$ is a prototype associated with the jth cluster; the k-means algorithm generates the clusters using the following objective function:

$$\min_{\{\psi_{ij}, c_j\}} \sum_{i=1}^{n} \sum_{j=1}^{k} \psi_{ij} \|s_i - c_j\|_2^2, \tag{4}$$

where ψ_{ij} is a Boolean indicator variable setting to either 1 when local descriptor s_i is assigned to cluster j or 0 otherwise. In VLAD, k-means algorithm transforms each local feature descriptor from the feature space to codeword by performing such hard assignment.

In FV, GMM is used to convert local feature set $S = \{s_1, s_2, \cdots, s_n\}$ to the codeword by performing soft assignment for each local feature s_i. Suppose that k clusters (i.e., Gaussian distributions each representing a visual word) are required, the distribution, represented by the probability density function (PDF) $\mathcal{P}$, over the entire feature space can be described as:

$$\mathcal{P}(s_i|\gamma) = \sum_{j=1}^{k} w_j \cdot \mathcal{N}(s_i|\mu_j, \sum_j^2), \tag{5}$$

where $\gamma = \{w_1, \mu_1, \sum_1^2, \cdots, w_k, \mu_k, \sum_k^2\}$, w_j is the weight for encoding relative frequency of visual word j subject to $\sum_{j=1}^{k} w_j = 1$, $\mathcal{N}(|)$ is a Gaussian distribution whose mean and covariance are μ_j and $\sum_j^2$ respectively regarding visual word j. Practically, the optimal model parameters γ can be learned via the maximum likelihood estimation (MLE) using the EM algorithm in an iterative manner. Theoretically, GMM is more powerful than k-means during the process of codebook generation as it is associated with both the mean information of the code words and the shape of their distributions. However, the performance for a given case depends practically on the sparseness of the features and the number of utilized Gaussian distributions.

3) Post-processing

In order to enable the encoded features $\hat{X}$ (in column-wise) to be invariant to the number of extracted local descriptors, they are usually normalized using $\ell 1$, $\ell 2$, power normalization (PN), or intra-normalization (IN) [32]. In particular, PN plus $\ell 2$-normalization (PNL2) [32] is applied in this work to reduce the number of outliers (or peaks) within the encoded feature vectors:

$$\hat{X}_i^{PNL2} = \|\text{sign}(\hat{X}_i)|\hat{X}_i|^{\alpha}\|_{\ell_2}, \tag{6}$$

where $\alpha \in [0, 1]$. After normalization, all the generated features can be readily fed into a classifier for action recognition. A good number of classifiers have been well studied and discussed in the literature for video action recognition purpose, and thus the discussion of this phase is omitted here.

IV. EAR-BASED MEMORY SUPPORT SYSTEM

The proposed gaze-informed egocentric action recognition approach has been applied to a memory support system for people with poor memory such as dementia, which illustrates the effectiveness of the proposed egocentric action recognition system in real-world application.

A. SYSTEM OVERVIEW

The working flow of the memory aid system is shown below in Fig. 3. The input of the system is an egocentric video stream integrated with gaze information which recorded all the activities (or events) happened to and witnessed by the patient for a certain period of time. This stream is then segmented into a set of video clips and each of which corresponds to an independent activity or event. Note that a sudden change of color, motion and other visual information in consecutive frames will occur if there is an activity change, which can be practically utilized to implement the temporal segmentation task. Multiple implementations are available for such operations, such as pair-wise pixel comparison,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808486, IEEE Access

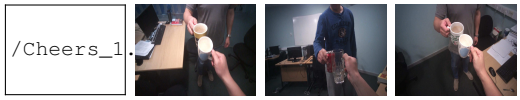
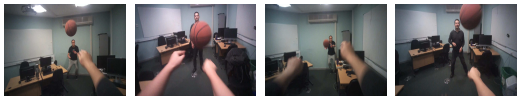
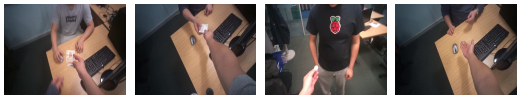
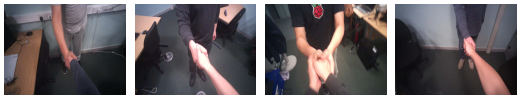
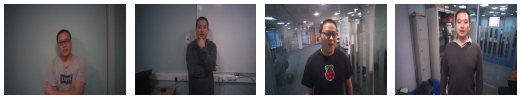
block-based comparison, and histogram comparison [33], and the detailed implementations are omitted here.

The segmented videos are classified to a certain number of ADLs by applying the gaze-informed egocentric action recognition approach, as introduced in section III. The patients or carers will then be able to search for a specific event by typing or speaking some keywords. The memory aid system then returns the video clips that mostly match the user specified query, and the patients and carers can watch the video clips for decision making or further actions. Note that the general purpose training data set may not be perfectly suitable for a particular user, and thus the returned results may not be sufficiently accurate. In this case, the patients or carers will have the opportunity to correct the labels of the returned video clips and to include these misclassified ones in the training data set for better coverage. Of course, the classification model will be trained again in this situation and after every such training data set updating event. Thanks to this feedback loop, the system is able to provide robust performance in an adaptive manner.

B. TRAINING DATA SET

An initial training data set is required to enable the proposed memory aid system to perform. In this work, the initial training data set was collected using the Tobii Pro® Glasses 2, named as UNN-GazeEAR ADL data set. This data set contains 50 human-to-human interaction video clips in 7586 frames in total. Each video clip is of the same resolution of 1920-by-1080 pixels with a unified frame rate of 25 frames per second (FPS) but with different time duration ranging from 2 to 11 seconds. The data set is associated with frame-wise gaze point which is determined based on a built-in eye fixation analysis function. The video clips represent five categories (each with 10 videos) of common human to human interactive ADLs as illustrated in Table 1, including 'Greeting', 'Passing a Ball', 'Paying', 'Shaking Hands', and 'Talking'. The video clips were captured under different conditions, such as diverse lighting conditions, various head postures from the wearer, as well as fast and frequently changed viewpoints with dynamic backgrounds. The building of a data set with wider coverage remains as active future work.

TABLE 1: Illustrative frames of UNN-GazeEAR dataset

Category	Sample Frames
Greeting	
Passing a Ball	
Paying	
Shaking Hands	
Talking	

V. EXPERIMENTATION

All the experiments were carried out using a HP workstation with Intel® Xeon™ E5-1630 v4 CPU @ 3.70 GHz. For fast processing, a sampling rate of 6 frames along the time line during the feature extraction process was applied in all the experiments. In the post-processing phase, the normalization parameter α in PNL2 was set as 0.5. The back-propagation neural network (BPNN) was employed in this work for ADL classification by taking the encoded and normalized feature values as inputs. The BPNN was trained using the scaled conjugate gradient (SCG) method [34] and the training ratio is 70%. The number of hidden neurons in the BPNN was uniformly set as 20. The performance was measured using the mean average precision, shorten as precision or accuracy hereafter, over 100 independent runs. Four groups of experiments are reported in this section in evaluating: 1) the effectiveness of GROI, 2) the impact of the resolution and block sizes, 3) the impact of the feature dimensions, and 4) the overall performance of the memory support system.

A. THE EFFECTIVENESS OF GROI

The effectiveness of GROI was evaluated by comparing the performance of the proposed gaze-informed egocentric action recognition approach and the conventional approach based on all the content in the video frames, using the UNN-GazeEAR data set in both sparse and dense representations, as shown in Table 2. In this experiment, the original resolution of 1920-by-1080 pixels was kept for all the videos; 72 dimensions were used after the PCA operation in the pre-processing phase; and the number of centroids in VLAD and FV were valued as 64 and 32 respectively; different block sizes, from 16-by-16 spatial pixels by 6 temporal frames to 64-by-64 pixels by 6 frames, were also investigated.

The performances led by the proposed and the conventional action recognition approaches with different parameters are listed in Table 2. The best results of both approaches under different feature extraction and encoding approaches are marked in bold. Thanks to the background removal ability, after the application of GROI identification operation, less storage memory space is required, which significantly reduced the action recognition processing time as shown in the table. It is also clear from the table that the proposed approach outperforms the conventional one without the use of GROI, which reveals the high degree of spatio-temporal variations in visual appearance (i.e., viewpoint changes, intra-class variation, and camera motion) of the UNN-GazeEAR ADL dataset, but also the power of GROI in action recognition from egocentric video clips.

It is interesting that smaller block sizes generally led to

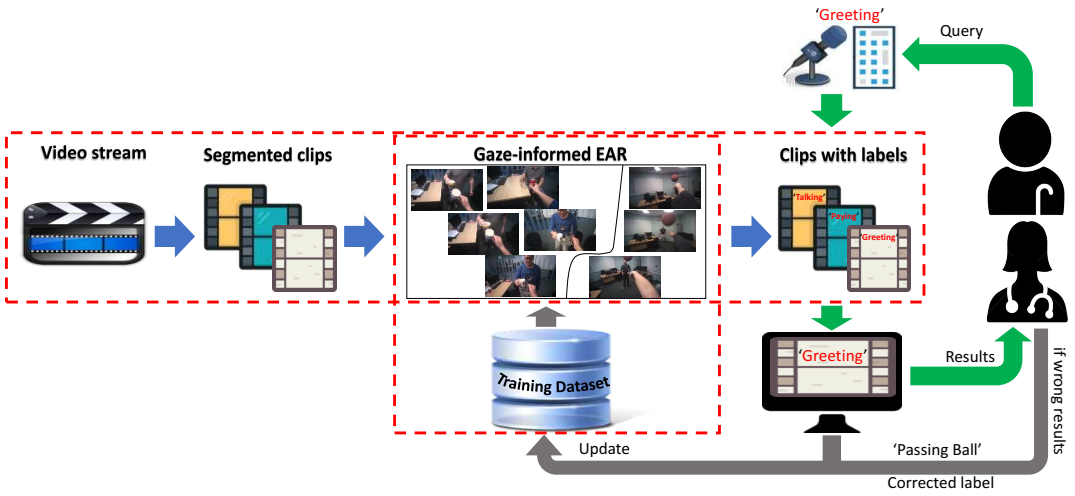


FIGURE 3: The proposed memory aid system

TABLE 2: Performance comparison of gaze-informed EAR and the conventional approach without gaze and ROI information under different conditions (the elapsed time of feature extraction is measured in seconds)

FeatDesc	UNN-GazeEAR without GROI @1920 × 1080 (826 MB)									UNN-GazeEAR with GROI @1920 × 1080 (521 MB)								
	[16 × 16 × 6]			[32 × 32 × 6]			[64 × 64 × 6]			[16 × 16 × 6]			[32 × 32 × 6]			[64 × 64 × 6]		
	Time	*VLAD*	*FV*	*Time*	*VLAD*	*FV*	*Time*	*VLAD*	*FV*	*Time*	*VLAD*	*FV*	*Time*	*VLAD*	*FV*	*Time*	*VLAD*	*FV*
HOG	1433	87.34	**86.78**	1528	**88.38**	86.42	1388	87.48	86.66	865	**95.60**	**97.28**	817	94.40	95.18	802	93.94	93.74
HMG	1419	88.86	87.68	1410	88.88	**88.14**	1383	**89.68**	87.10	859	**95.94**	**95.82**	815	94.68	92.52	802	93.64	93.62
HOF	1644	**87.06**	**86.72**	1651	86.96	86.26	1635	85.86	86.10	1077	94.58	**95.30**	1021	**96.00**	94.28	1011	95.02	93.64
MBHx	2216	86.68	85.56	2198	87.50	**86.16**	2675	**88.68**	85.32	1767	92.78	**95.82**	1579	93.20	95.28	1559	**94.50**	95.08
MBHy	2233	86.42	84.34	2253	85.58	84.54	2180	**88.42**	**85.44**	1764	**97.44**	**94.72**	1575	94.92	93.28	1558	94.10	93.98

much better results in this experiment for the proposed approach, but this was not the case for the conventional method without the use of GROI. In particular, HOG, HMG, and MBHy achieved their best performances using the smallest block size [16 × 16 × 6], while HOF and MBHx reached their peak performances under the FV encoding scheme. The proposed approach reached its best performance of 97.44% using the feature extraction MBHy (with the block size of [16 × 16 × 6]) and feature encoding approach VLAD, while the conventional action recognition approach achieved its peak performance of 89.68% when HMG and VLAD with a block size of [64 × 64 × 6] were used.

This experiment demonstrates the effectiveness of GROI in action recognition. The conventional approach may suffer from moving background and other irrelevant foreground activities, but the proposed approach mitigates such limitation by removing the noise before the processing stage. Also, the proposed approach is proved to be a computationally efficient method, which computes based only on the most relevant information in the egocentric video clips.

B. THE IMPACT OF RESOLUTION AND BLOCK SIZE

Generally speaking, higher resolution and smaller block size lead to higher computational cost, and vice verse. Note that the proposed memory support system may be required to perform online in some situations; thus the resolution and block sizes were empirically studied in this experiment. In particular, the resolution of the UNN-GazeEAR data set was down-scaled with a scaling factor of 6 (i.e., $\Omega = 6$) and the performance based on the down-scaled data set is demonstrated in Table 3. To facilitate the comparative study, the same parameter values that used for the last experiment as reported in Sub-section V-A were also used in this experiment. Note that the block size of [64 × 64 × 6] was not included in this experiment because the block size was too big for the down-scaled resolution of 320 × 180 videos, and this was also the case for block size of [32 × 32 × 6] under the FV encoding scheme.

The results shown in Table 3 based on resolution 320-by-180 suggest that the feature extraction speed could be significantly boosted up by around 50 times without obvious sacrifice of accuracy, in reference to the performance based on the original resolution of 1920-by-1080 pixels using the same gaze-informed EAR approach. This experiment also indicates that the feature extraction speed can also be boosted up by using a smaller block size, again without obvious performance deterioration. Among the five applied feature extraction approaches, HOG and HMG were not only the fastest with exactly the same time consumption, but also overall the best. In specific, 22, 18, 17, and 16 seconds were used regarding the four block sizes from small to large, respectively. This experiment demonstrates that the proposed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808486, IEEE Access

TABLE 3: The mAP (%) of family of local spatio-temporal descriptors with 72 feature dimensions under both encoding schemes by varying the block size using the UNN-GazeEAR data set that down-scaled by a factor of 6.

FeatDesc	UNN-GazeEAR with GROI @320 × 180 (39.1 MB)											
	[4 × 4 × 6]			[8 × 8 × 6]			[16 × 16 × 6]			[32 × 32 × 6]		
	Time	*VLAD*	*FV*	*Time*	*VLAD*	*FV*	*Time*	*VLAD*	*FV*	*Time*	*VLAD*	*FV*
HOG	22	96.34	96.48	18	94.62	94.56	17	95.00	93.96	16	94.50	93.80
HMG	22	95.40	96.08	18	93.82	92.60	17	94.28	92.28	16	95.74	–
HOF	26	93.94	91.22	22	93.90	92.94	21	95.62	91.98	20	92.88	–
MBHx	41	90.72	91.80	33	93.14	92.48	32	95.88	92.12	30	94.34	–
MBHy	41	93.74	91.46	33	94.66	91.20	32	95.78	93.54	30	95.20	–

gaze-informed EAR has good scale-invariance property in terms of video resolution and block size.

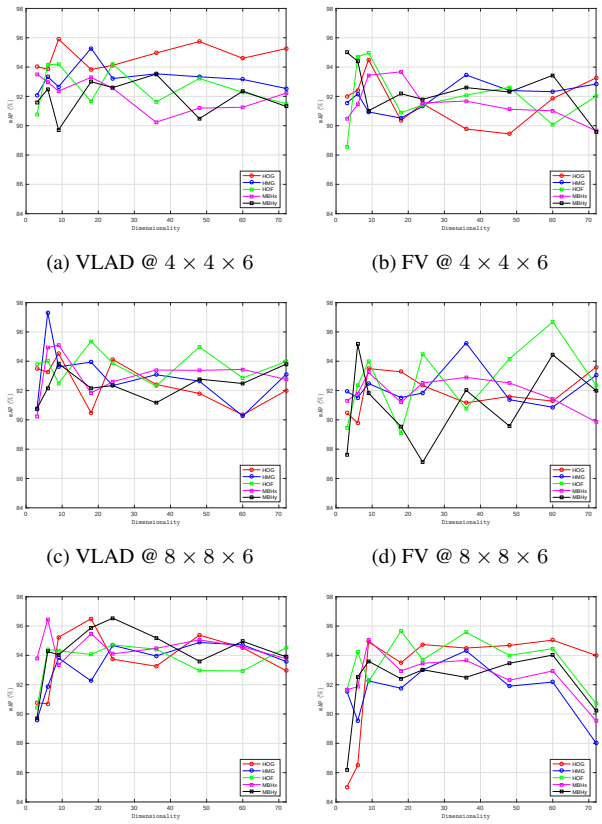
C. THE IMPACT OF FEATURE DIMENSIONS AND ENCODING APPROACHES

This experiment investigated the effect of the dimensionality of the extracted features in the pre-processing phase using the PCA. In this experiment, the down-scaled data set was used and the number of centroids for VLAD and FV were set to 32 and 16, respectively for fast processing. The experimental results using 3, 6, 9, 18, 24, 36, 48, 60 and 72 dimensions with different feature extraction and encoding approaches are demonstrated in Fig. 4. It is difficult to observe any obvious performance variation based on the studied feature dimensions except that FV performed very poorly when low feature dimensions were used, which demonstrates the robustness of the proposed approach.

The best performances of the proposed system using different feature dimensions with various feature extraction approaches and block sizes are summarized in Table 4. The overall optimal performance was achieved at dimension value of 9 when block size $[4 \times 4 \times 6]$ was used (i.e., 95.90% achieved by HOG), at dimension value of 6 when $[8 \times 8 \times 6]$ was used (i.e., 97.32% yielded by HMG), and at dimension value of 24 when $[16 \times 16 \times 6]$ was used (i.e., 96.54% generated using MBHy). When it comes to the five applied feature descriptors, none of these achieved their best accuracy using the smallest block size $[4 \times 4 \times 6]$ in this experiment. HMG obtained its highest accuracies 97.32% with 6 feature dimensions under VLAD, and HOF achieved it optimal performance 96.68% with 60 feature dimensions under FV, using the block size of $[8 \times 8 \times 6]$. HOG achieved its best accuracies 96.50% with 18 feature dimensions under VLAD, and MBHx reached its peak performances 96.46% with 6 dimensions, and MBHy achieved its best performance 96.54% with 24 feature dimensions under VLAD, using the block size of $[16 \times 16 \times 6]$.

D. THE MEMORY AID SYSTEM

This experiment examined the proposed memory support system using an untrimmed video stream that simulated the five ADLs studied in this paper. In particular, the optimal fine-tuned parameters discovered from the experiments as reported in the previous subsections were utilized in the integrated gaze-informed egocentric action recognition system. The video stream has 915 frames and it includes five continuously happened actions, which were particularly arranged and recorded in this way to support this investigation. The video stream was firstly segmented into 5 video clips by the memory support system as shown in the top of Table 5, representing 'Passing a Ball' in frames 1-100, 'Shaking Hands' in frames 101-194, 'Talking' in frames 195-340, 'Paying' in frames 341-648, and 'Greeting' in frames 649-915; the details of segmentation is omitted here as this is beyond the focus of this paper.



(a) VLAD @ $4 \times 4 \times 6$ (b) FV @ $4 \times 4 \times 6$

(c) VLAD @ $8 \times 8 \times 6$ (d) FV @ $8 \times 8 \times 6$

(e) VLAD @ $16 \times 16 \times 6$ (f) FV @ $16 \times 16 \times 6$

FIGURE 4: Performance comparison between local features with different feature dimensions

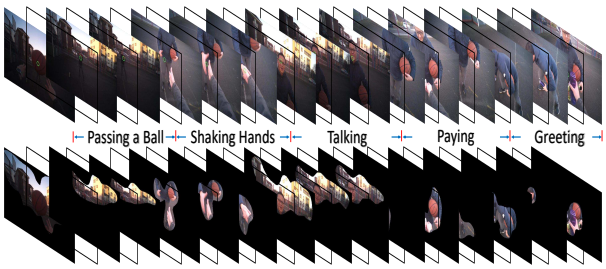
The proposed gaze-informed egocentric action recogni-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2808486, IEEE Access

TABLE 4: Accuracy comparison using different feature dimensions with various block sizes (the adopted color schemes is consist with Fig. 4)

Block Size	UNN-GazeEAR with GROI @320 × 180 (39.1 MB)																	
	3		*6*		*9*		*18*		*24*		*36*		*48*		*60*		*72*	
	VLAD	*FV*	*VLAD*	*FV*	*VLAD*	*FV*	*VLAD*	*FV*	*VLAD*	*FV*	*VLAD*	*FV*	*VLAD*	*FV*	*VLAD*	*FV*	*VLAD*	*FV*
$[4 \times 4 \times 6]$	*94.02*	95.00	*94.16*	94.70	***95.90***	94.96	*95.28*	93.68	*94.20*	91.80	*94.96*	93.46	*95.74*	92.62	*94.60*	93.44	*95.26*	93.26
$[8 \times 8 \times 6]$	*93.80*	91.94	***97.32***	95.16	*95.10*	94.00	*95.36*	93.28	*94.12*	94.50	*93.40*	95.24	*94.98*	94.16	*93.44*	96.68	*93.98*	93.60
$[16 \times 16 \times 6]$	*93.78*	91.66	*96.46*	94.22	*95.24*	95.06	*96.50*	95.66	***96.54***	94.74	*95.20*	95.58	*95.38*	94.68	*94.96*	95.04	*94.52*	94.00

TABLE 5: Testing video stream for the memory aid system with five ADLs



HOG (VLAD)	✓	✗	✓	✓	✓
HOG (FV)	✓	✓	✓	✓	✓
HMG (VLAD)	✓	✗	✓	✗	✓
HMG (FV)	✓	✓	✓	✗	✓
HOF (VLAD)	✓	✗	✓	✗	✓
HOF (FV)	✓	✓	✓	✗	✓
MBHx (VLAD)	✓	✓	✓	✗	✓
MBHx (FV)	✓	✗	✓	✓	✗
MBHy (VLAD)	✓	✓	✓	✗	✓
MBHy (FV)	✓	✓	✓	✗	✓

tion was applied to the segmented video clips for labelling, which are summarized in Table 5. From this table, HOG under FV encoding scheme achieved the best performance (100%) which expressed the good generalization ability of such a feature generation pipeline. Interestingly, the action clip themed with 'Paying' (with the largest number of frames compared with the rest four) has been wrongly predicted as 'Passing Ball' or 'Greeting' using all the feature descriptors except HOG. This might be caused by the high frequency of appearances of the irrelevant background objects basketball and drinking cups, which appeared in ten 'Passing Ball' and 'Greeting' training video clips in the UNN-GazeEAR training data set but not the ten 'Paying' clips. In this situation, the performance of the system can be improved by adding the misclassified clips with the right labels into the training data set. In other words, the human-in-the-loop control integrated in the proposed memory aid system provides not only good system adaptability but also human-centered functionality.

E. DISCUSSIONS

The performance of the proposed gaze-informed egocentric action recognition system in the above experiments validated the working of the system, and proved the computational efficiency (in terms of both space and time complexity) in recognizing visual egocentric actions. The investigation on the memory aid system demonstrated its effectiveness, but the usability of the system requires further study which remains as future work.

VI. CONCLUSIONS

This paper proposed a gaze-informed egocentric action recognition approach, which utilizes gaze information to inform the saliency-based ROI determination for background and noise removal. The proposed approach has been further extended to a memory aid system to support people with poor memory. The proposed approach was evaluated and analyzed using a in-house captured egocentric ADL data set UNN-GazeEAR. The experimental results demonstrate the functional efficacy and computational efficiency of the proposed memory aid system and its underpinning egocentric action recognition approach. Noting that only several actions were empirically studied in this work, it is therefore important in the future to extend the data set to cover a wider range of action categories and eventually to generalize the memory aid system. Also, a large data set is required to intensively evaluate the saliency ROI detection and the action recognition, in addition to the verification of the feature representation algorithms presented in this work. What is more, the proposed GROI method can be extended to other visual recognition tasks such as real-time egocentric visual navigation for blind people and mobile robots.

REFERENCES

[1] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PP, no. 99, pp. 1–1, 2017.

[2] M. A. Uddin, J. B. Joolee, A. Alam, and Y. K. Lee, "Human action recognition using adaptive local motion descriptor in spark," IEEE Access, vol. 5, pp. 21 157–21 167, 2017.

[3] T. Wang, Y. Chen, M. Zhang, J. Chen, and H. Snoussi, "Internal transfer learning for improving performance in human action recognition for small datasets," IEEE Access, vol. 5, pp. 17 627–17 633, 2017.

[4] M. S. Ryoo and L. Matthies, "First-person activity recognition: Feature, temporal structure, and prediction," International Journal of Computer Vision, vol. 119, no. 3, pp. 307–328, 2016.

[5] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in CVPR 2011, June 2011, pp. 3241–3248.

[6] E. H. Spriggs, F. D. L. Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, June 2009, pp. 17–24.
[7] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1894–1903.
[8] T.-H.-C. Nguyen, J.-C. Nebel, and F. Florez-Revuelta, "Recognition of activities of daily living with egocentric vision: A review," Sensors, vol. 16, no. 1, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/1/72
[9] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On. IEEE, 2011, pp. 3281–3288.
[10] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," IEEE transactions on pattern analysis and machine intelligence, vol. 34, no. 1, pp. 194–201, 2012.
[11] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki, Can Saliency Map Models Predict Human Egocentric Visual Attention? Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 420–429.
[12] Z. Wang and J. Liu, "A review of object detection based on convolutional neural network," in 2017 36th Chinese Control Conference (CCC), July 2017, pp. 11 104–11 109.
[13] I. Laptev, "On space-time interest points," Int. J. Comput. Vision, vol. 64, no. 2-3, pp. 107–123, Sep. 2005.
[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Computer Vision and Pattern Recognition, 2005., vol. 1. IEEE, 2005, pp. 886–893.
[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in CVPR'08, June 2008, pp. 1–8.
[16] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in ECCV'06, 2006, pp. 428–441.
[17] I. C. Duta, J. R. R. Uijlings, T. A. Nguyen, K. Aizawa, A. G. Hauptmann, B. Ionescu, and N. Sebe, "Histograms of motion gradients for real-time video classification," in Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on. IEEE, 2016, pp. 1–6.
[18] R. Cameron, Z. Zuo, G. Sexton, and L. Yang, "A fall detection/recognition system and an empirical study of gradient-based feature extraction approaches," in UK Workshop on Computational Intelligence. Springer, 2017, pp. 276–289.
[19] H. Wang and C. Schmid, "Action recognition with improved trajectories," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.
[20] A. Coates and A. Y. Ng, "Learning feature representations with k-means," in Neural networks: Tricks of the trade. Springer, 2012, pp. 561–580.
[21] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," Computer Vision and Image Understanding, vol. 150, pp. 109 – 125, 2016.
[22] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 23, no. 3, pp. 412–424, March 2012.
[23] S. Yan, X. Xu, D. Xu, S. Lin, and X. Li, "Beyond spatial pyramids: A new feature extraction framework with dense spatial sampling for image classification," Computer Vision–ECCV 2012, pp. 473–487, 2012.
[24] J. N. Kather, A. Weidner, U. Attenberger, Y. Bukschat, C.-A. Weis, M. Weis, L. R. Schad, and F. G. Zöllner, "Color-coded visualization of magnetic resonance imaging multiparametric maps," Scientific reports, vol. 7, 2017.
[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vision, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[26] J. Uijlings, I. C. Duta, E. Sangineto, and N. Sebe, "Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off," International Journal of Multimedia Information Retrieval, vol. 4, no. 1, pp. 33–44, 2015.
[27] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, "Real-time visual concept classification," IEEE Transactions on Multimedia, vol. 12, no. 7, pp. 665–681, Nov 2010.
[28] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 2911–2918.
[29] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: a comparative," J Mach Learn Res, vol. 10, pp. 66–71, 2009.
[30] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," IEEE transactions on pattern analysis and machine intelligence, vol. 34, no. 9, pp. 1704–1716, 2012.
[31] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007, pp. 1–8.
[32] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in ECCV'10, 2010, pp. 143–156.
[33] I. Koprinska and S. Carrato, "Temporal video segmentation: A survey," Signal processing: Image communication, vol. 16, no. 5, pp. 477–500, 2001.
[34] C. Xia, Z. Yang, B. Lei, and Q. Zhou, "Scg and lm improved bp neural network load forecasting and programming network parameter settings and data preprocessing," in 2012 International Conference on Computer Science and Service System, Aug 2012, pp. 38–42.

Zheming Zuo received the B.Eng. degree in computer science and technology from the Southeast University, Nanjing, China, and the M.Sc. degree in computer science from the University of Birmingham, Birmingham, U.K., respectively, in 2012 and 2014.

He is currently a Ph.D. candidate with research interests include computer vision, machine learning, and fuzzy logic systems.

Longzhi Yang is currently a Programme Leader and Senior Lecturer with Northumbria University, Newcastle, U.K. His research interests include computational intelligence, machine learning, big data, computer vision, intelligent control systems, and the application of such techniques in real-world uncertain environments. He is the founding chair of IEEE Special Interest Group (SIG) on Big Data for Cyber Security and Privacy.

Dr. Yang received the Best Student Paper Award at the 2010 IEEE International Conference on Fuzzy Systems.

Yonghong Peng is a Professor of Data Science and the leader for Data Science Research at the University of Sunderland, United Kingdom. His research areas include Data Science, Machine Learning, Data Mining and Artificial Intelligence. He is the Chair for the Big Data Task Force (BDTF), and a member of Data Mining and Big Data Analytics Technical Committee of IEEE computational intelligence society (CIS). He is also a founding member of the Technical Committee on Big Data (TCBD) of IEEE Communications and an advisory board member for IEEE Special Interest Group (SIG) on Big Data for Cyber Security and Privacy. Prof Peng is an Associate Editor for IEEE Transaction on Big Data, and an Academic Editor of PeerJ and PeerJ Computer Science.

Fei Chao received the B.Sc. degree in mechanical engineering from Fuzhou University, Fuzhou, China, and the M.Sc. degree with distinction in computer science from the University of Wales, Cardiff, U.K., in 2004 and 2005, respectively, and the Ph.D. degree in robotics from Aberystwyth University, Aberystwyth, U.K., in 2009ãĂĆ

He was a Research Associate under the supervision of Prof. M. H. Lee with Aberystwyth University from 2009 to 2010. He is currently an Associate Professor with the Cognitive Science Department, Xiamen University, Xiamen, China. He has published 30 peer-reviewed journal and conference papers. His research interests include developmental robotics, machine learning, and optimization algorithms.

Dr. Chao is the Vice-Chair of the IEEE Computer Intelligence Society Xiamen Chapter. He is also a member of CCF.

Yanpeng Qu received a Ph.D. degree in Computational Mathematics from Dalian University of Technology, China. He is an Associate Professor with the Information Science and Technology College at Dalian Maritime University, China. His current research interests include granular computing, neural networks, pattern recognition, data mining, and real-world applications of such techniques for intelligent decision support.